Kidney Disease PubMed

Janis Corona

12/11/2019

This script takes articles from the abstracts on Kidney Disease articles from NCBI's PubMed, PLOS, and the summary of the NCBI GEO sample pages

This creates a directory to stem the abstracts and preprocess from the csv file into a corpus of 20 files in a folder called KidneyDisease.

```r
Auto <- read.csv('NIH_PLOS_articles_kidney_disease.csv', sep=',',
                 header=FALSE, na.strings=c('',' '))
```

```r
colnames(Auto) <- c('abstract','source')
auto <- Auto[complete.cases(Auto$abstract),]


dir.create('./KidneyDisease')

ea <- as.character(auto$abstract)
setwd('./KidneyDisease')

for (j in 1:length(ea)){
  write(ea[j], paste(paste('EA',j, sep='.'), '.txt', sep=''))
}
setwd('../')
```

This code preprocesses and stems the corpus

```r
library(tm)
library(SnowballC)
library(wordcloud)
library(ggplot2)

KidneyDisease <- Corpus(DirSource("KidneyDisease"))


KidneyDisease
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 43
```

```r
#KidneyDisease <- tm_map(KidneyDisease, removePunctuation)
#KidneyDisease <- tm_map(KidneyDisease, removeNumbers)
KidneyDisease <- tm_map(KidneyDisease, tolower)
KidneyDisease <- tm_map(KidneyDisease, removeWords, stopwords("english"))
KidneyDisease <- tm_map(KidneyDisease, stripWhitespace)
KidneyDisease <- tm_map(KidneyDisease, stemDocument)
```

```
dtmKidneyDisease <- DocumentTermMatrix(KidneyDisease)

freq <- colSums(as.matrix(dtmKidneyDisease))
```

This code orders words stemmed by frequency and finds input correlations

```
FREQ <- data.frame(freq)
ord <- order(freq, decreasing=TRUE)

freq[head(ord, 25)]
```

```
##     kidney     medium     associ       cell      serum supplement
##        223        128        112        110        102         98
##     sodium   concentr     diseas    univers       egfr     depart
##         97         82         77         77         75         74
##     declin        use    purchas   function      renal  medicine,
##         71         68         64         64         63         61
##      sampl       risk     growth      incid      tissu      rapid
##         57         54         51         51         50         50
##        san
##         50
```

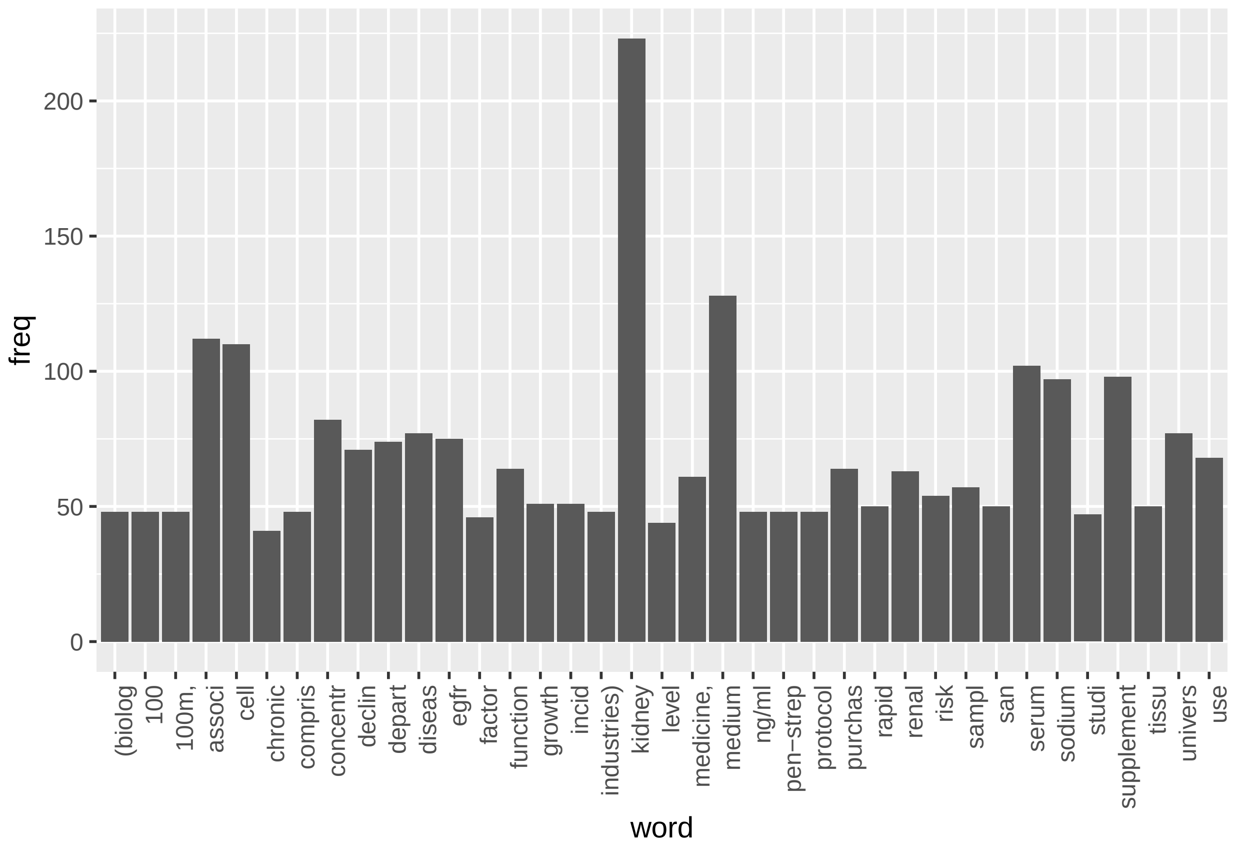
```
findAssocs(dtmKidneyDisease, "renal", corlimit=0.5)
```

```
## $renal
##             mice           calcul              (b)              .e.
##             0.70             0.69             0.68             0.68
##       accomplish             area            area.        ascertain
##             0.68             0.68             0.68             0.68
##              axi            axial            axis,          biochem
##             0.68             0.68             0.68             0.68
##          biopsy.              can        ckd-relat         collagen
##             0.68             0.68             0.68             0.68
##         content,         content.            coron          deposit
##             0.68             0.68             0.68             0.68
##        distance,           easili           ellips        ellipsoid
##             0.68             0.68             0.68             0.68
##           extend           extent          formula             imag
##             0.68             0.68             0.68             0.68
##       interstiti        invasive,             just             make
##             0.68             0.68             0.68             0.68
##            minor         noninvas              now            often
##             0.68             0.68             0.68             0.68
##           organ.        parenchym            pelvi      picrosirius
##             0.68             0.68             0.68             0.68
##            polar              red           remark           risky,
##             0.68             0.68             0.68             0.68
##             scar        scarring,             size            size,
##             0.68             0.68             0.68             0.68
##          sometim            stain         techniqu            today
##             0.68             0.68             0.68             0.68
##            treat             true tubulointerstiti      ultrasound,
##             0.68             0.68             0.68             0.68
##       underestim              via           visual            major
##             0.68             0.68             0.68             0.65
##           obtain           involv
##             0.52             0.51
```

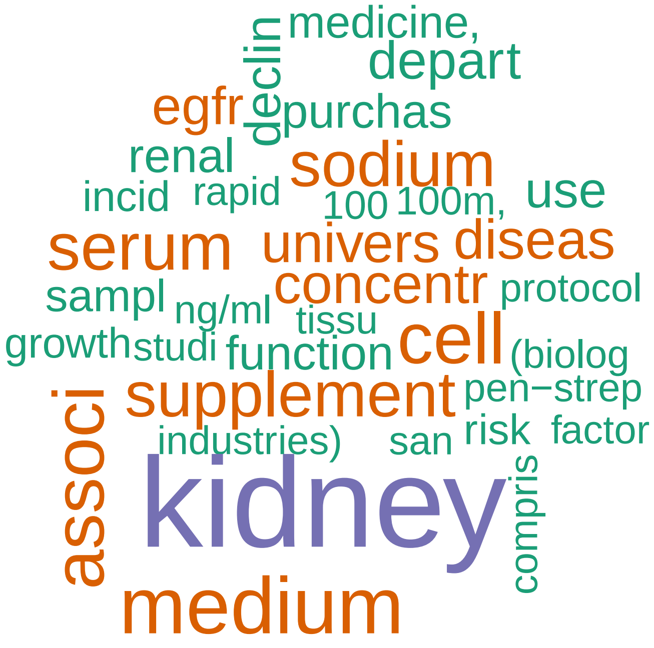
```
findAssocs(dtmKidneyDisease, "pain", corlimit=0.69)
```

```
## $pain
##             (12)       (23.0-28.0            (25%)            (92%)
##              0.7              0.7              0.7              0.7
##          (health  (sd)ml/min/1.73       0.13-0.97)       0.97-3.07)
##              0.7              0.7              0.7              0.7
##            1.72;              1.9           2-fold             25.2
##              0.7              0.7              0.7              0.7
##              252            70-79              989            [8%])
##              0.7              0.7              0.7              0.7
##             ab9,              abc             abc)           aging,
##              0.7              0.7              0.7              0.7
##           aging;       alkalosis;        analyzer.    anesthesiolog
##              0.7              0.7              0.7              0.7
##        bethesda,       california            city,         composit
##              0.7              0.7              0.7              0.7
##             de4,         egfr0.55          elders:           forest
##              0.7              0.7              0.7              0.7
##              fri             give            harri       inception.
##              0.7              0.7              0.7              0.7
##          insight         interven   investigators.            jh12;
##              0.7              0.7              0.7              0.7
##      kritchevski             kv5,             lake            least
##              0.7              0.7              0.7              0.7
##             lf3,             lost            m(2),            mg11,
##              0.7              0.7              0.7              0.7
##            mj10,           mmol/l         mmol/l),          mmol/l.
##              0.7              0.7              0.7              0.7
##           newman              pa.              pa;            patel
##              0.7              0.7              0.7              0.7
##         persons.      pittsburgh,         predomin     progression,
##              0.7              0.7              0.7              0.7
##           ratio.             rh6,           rifkin             salt
##              0.7              0.7              0.7              0.7
##             sb8,            separ           sticht             tb7,
##              0.7              0.7              0.7              0.7
##             th2,              ut.            utah,           venous
##              0.7              0.7              0.7              0.7
##             wake       well-funct   winston-salem,          yenchek
##              0.7              0.7              0.7              0.7
##              (b)              .e.       accomplish             area
##              0.7              0.7              0.7              0.7
##            area.        ascertain              axi            axial
##              0.7              0.7              0.7              0.7
##            axis,          biochem          biopsy.              can
##              0.7              0.7              0.7              0.7
##        ckd-relat         collagen         content,         content.
##              0.7              0.7              0.7              0.7
##            coron          deposit        distance,           easili
##              0.7              0.7              0.7              0.7
##           ellips        ellipsoid           extend           extent
##              0.7              0.7              0.7              0.7
##          formula             imag       interstiti        invasive,
##              0.7              0.7              0.7              0.7
##             just             make            minor         noninvas
##              0.7              0.7              0.7              0.7
##              now            often           organ.        parenchym
##              0.7              0.7              0.7              0.7
##            pelvi      picrosirius            polar              red
##              0.7              0.7              0.7              0.7
##           remark           risky,             scar        scarring,
##              0.7              0.7              0.7              0.7
##             size            size,          sometim            stain
##              0.7              0.7              0.7              0.7
##         techniqu            today            treat             true
##              0.7              0.7              0.7              0.7
## tubulointerstiti      ultrasound,       underestim              via
##              0.7              0.7              0.7              0.7
##           visual
##              0.7
```

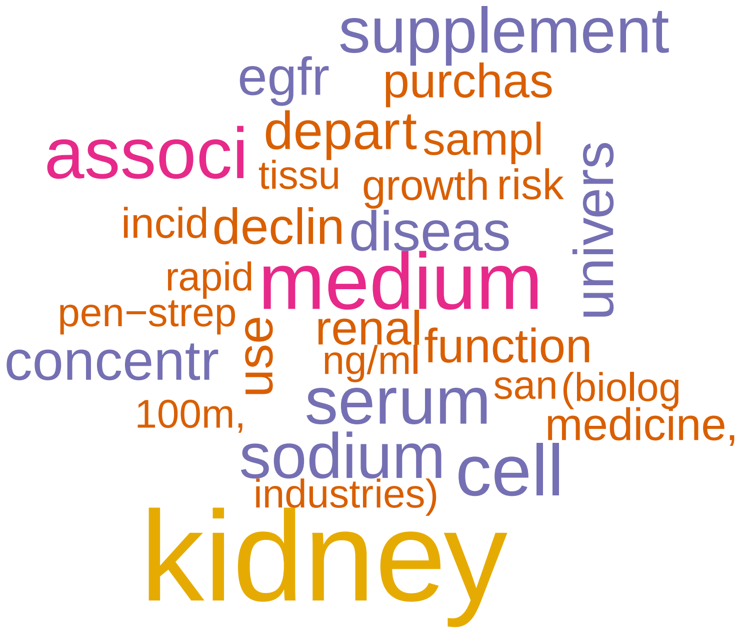
```r
wf <- data.frame(word=names(freq), freq=freq)
p <- ggplot(subset(wf, freq>40), aes(word, freq))
p <- p + geom_bar(stat= 'identity')
p <- p + theme(axis.text.x=element_text(angle=90, hjust=1))
p
```



```r
wordcloud(names(freq), freq, min.freq=45,colors=brewer.pal(3,'Dark2'))
```



```r
wordcloud(names(freq), freq, max.words=30,colors=brewer.pal(6,'Dark2'))
```



The above stemmed the corpus, this will lemmatize the original csv file

and add the field to the table and write out to csv, followed by plot the word count frequencies that were lemmatized and the word clouds

```r
library(textstem)

lemma <- lemmatize_strings(auto$abstract, dictionary=lexicon::hash_lemmas)

Lemma <- as.data.frame(lemma)
Lemma <- cbind(Lemma, auto)

colnames(Lemma) <- c('lemmatizedAbstract','abstract', 'source')

write.csv(Lemma, 'LemmatizedKidneyDisease.csv', row.names=FALSE)
```

```r
dir.create('./KidneyDisease-Lemma')

ea <- as.character(Lemma$lemmatizedAbstract)
setwd('./KidneyDisease-Lemma')

for (j in 1:length(ea)){
  write(ea[j], paste(paste('EAL',j, sep='.'), '.txt', sep=''))
}
setwd('../')
```

```r
library(tm)
library(SnowballC)
library(wordcloud)
library(ggplot2)
```

```r
KidneyDisease <- Corpus(DirSource("KidneyDisease-Lemma"))

KidneyDisease
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 43
```

```r
#KidneyDisease <- tm_map(KidneyDisease, removePunctuation)
#KidneyDisease <- tm_map(KidneyDisease, removeNumbers)
KidneyDisease <- tm_map(KidneyDisease, tolower)
KidneyDisease <- tm_map(KidneyDisease, removeWords, stopwords("english"))
KidneyDisease <- tm_map(KidneyDisease, stripWhitespace)

dtmKidneyDisease <- DocumentTermMatrix(KidneyDisease)
dtmKidneyDisease
```

```
## <<DocumentTermMatrix (documents: 43, terms: 2418)>>
## Non-/sparse entries: 7417/96557
## Sparsity           : 93%
## Maximal term length: 116
## Weighting          : term frequency (tf)
```

```r
freq <- colSums(as.matrix(dtmKidneyDisease))

FREQ <- data.frame(freq)
ord <- order(freq, decreasing=TRUE)

freq[head(ord, 25)]
```

```
##     kidney       cell     medium      serum     sodium supplement
##        223        142        128        102         97         96
##       egfr        100 invitrogen    disease university department
##         93         80         80         78         77         74
##    decline        use   function        4mg    aldrich biological
##         71         67         65         64         64         64
##   industry       poly   purchase        sfm      sigma      renal
##         64         64         64         64         64         63
##  associate
##         62
```

```r
pain <- as.data.frame(findAssocs(dtmKidneyDisease, "pain", corlimit=0.99))

kidney <- as.data.frame(findAssocs(dtmKidneyDisease, "kidney", corlimit=0.65))
```

```r
treatment <- as.data.frame(findAssocs(dtmKidneyDisease, "treatment", corlimit=0.81))

pain
```

```
##                 pain
## 1.9                1
## 23.                1
## 25.2               1
## 252                1
## 28.                1
## 72;                1
## 989                1
## ab9,               1
## abc                1
## age;               1
## alkalosis;         1
## analyzer.          1
## anesthesiology     1
## arterial           1
## arterialized       1
## bethesda,          1
## california         1
## city,              1
## collaborator       1
## collection         1
## composition        1
## de4,               1
## egfr0.55           1
## elder              1
## elder:             1
## forest             1
## fry                1
## give               1
## harris             1
## inception.         1
## insight            1
## intervene          1
## investigator.      1
## jh12;              1
## kritchevsky        1
## kv5,               1
## lake               1
## less               1
## lf3,               1
## lose               1
## mg11,              1
## mj10,              1
## mmol               1
## newman             1
## pa.                1
## pa;                1
## patel              1
## person.            1
## pittsburgh,       1
## predominantly     1
## prevalent         1
## progression,      1
## ratio.            1
## rh6,              1
## rifkin            1
## salem,            1
## salt              1
## sb8,              1
## separate          1
## sticht            1
## tb7,              1
## th2,              1
## ut.               1
## utah,             1
## venous            1
## wake              1
## winston           1
## yenchek           1
```

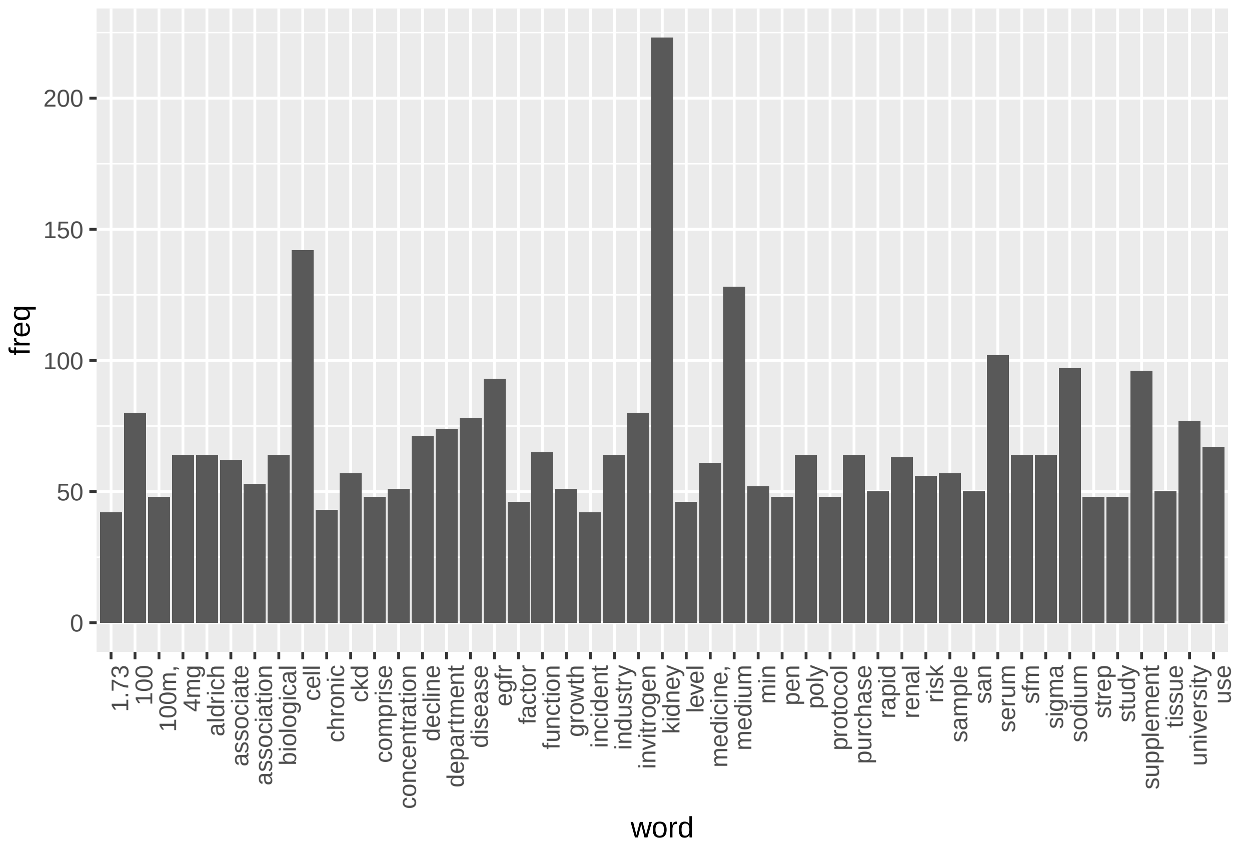
```
kidney
```

```
##              kidney
## function       0.72
## albuminuria    0.67
## ethnic         0.65
## katz           0.65
## washington,    0.65
```

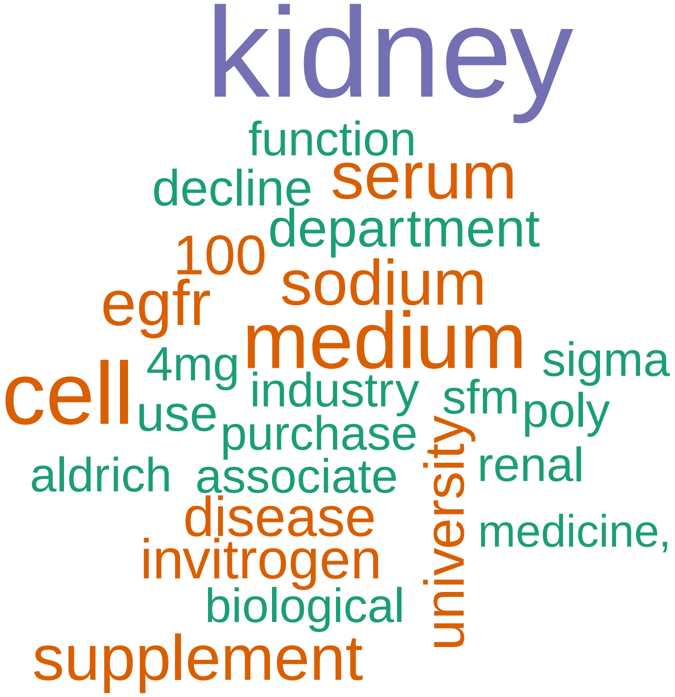
```
treatment
```

```
##       treatment
## cell       0.83
## lipid      0.83
## total      0.83
```

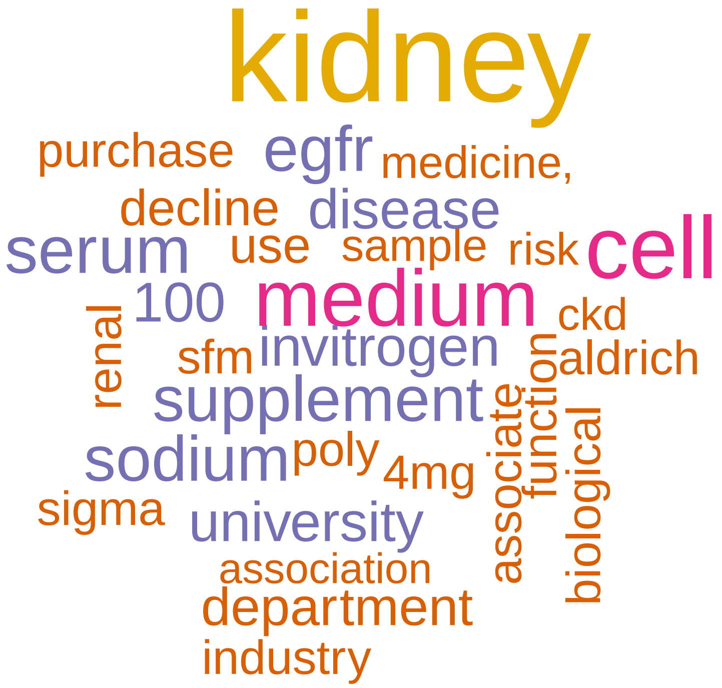
```
wf <- data.frame(word=names(freq), freq=freq)
p <- ggplot(subset(wf, freq>40), aes(word, freq))
p <- p + geom_bar(stat= 'identity')
p <- p + theme(axis.text.x=element_text(angle=90, hjust=1))
p
```



```
wordcloud(names(freq), freq, min.freq=60,colors=brewer.pal(3,'Dark2'))
```



```r
wordcloud(names(freq), freq, max.words=30,colors=brewer.pal(6,'Dark2'))
```


kidney
purchase
egfr
medicine,
decline
disease
serum
use
sample
risk
cell
100
medium
ckd
renal
sfm
invitrogen
function
aldrich
supplement
associate
biological
sodium
poly
4mg
sigma
university
association
department
industry